Partial Least Square Modeling for the Control of Refining Processes on Mid-Distillates by Near Infrared Spectroscopy

S. Aji, N. Schildknecht-Szydlowski, A. Faraj

Oil & Gas Science and Technology – Rev. IFP, Vol. 59 (2004), No. 3 pp. 303-321

Partial Least Square Modeling for the Control of Refining Processes on Mid-Distillates by Near Infrared Spectroscopy

S. Aji[1], N. Schildknecht-Szydlowski[1] and A. Faraj[1]

1 Institut français du pétrole, 1 et 4, avenue de Bois-Préau 92852 Rueil-Malmaison Cedex, France
e-mail: salaheddine.aji@ifp.fr - nathalie.schildknecht@ifp.fr - abdelaziz.faraj@ifp.fr

Résumé — Modèles PLS pour le suivi de la qualité par spectrométrie proche infrarouge des distillats moyens issus des procédés de raffinage — La régression PLS (*Partial Least Squares*) a été utilisée pour établir des modèles de prédiction de différentes familles chimiques *i.e.* % massique paraffines, naphtènes et % massique et mol/100 g monoaromatiques, diaromatiques+ et total aromatiques à partir de spectres proche infrarouge (PIR) de distillats moyens de composition chimique très variée. La classification a été utilisée pour tenir compte des ressemblances chimiques entre échantillons et organiser en classes la base de calibration. Les corrélations entre les spectres PIR et les propriétés modélisées permettent de prédire, pour la plupart, l'ensemble des propriétés dans deux fois l'intervalle de confiance à 95 % des méthodes de référence, après classification préalable des échantillons en trois *clusters*. La classification a été nécessaire pour améliorer la qualité de prédiction des modèles PLS.

Abstract* — *Partial Least Square Modeling for the Control of Refining Processes on Mid-Distillates by Near Infrared Spectroscopy — *Partial Least Squares regression (PLS) was used to elaborate the prediction models of the different chemical families i.e. wt% paraffins, naphthenes, and wt% and mol/100 g monoaromatics, diaromatics+ and total aromatics from Near InfraRed spectra (NIR) of mid-distillates covering a large range of chemical compositions. Cluster analysis was used to reveal the chemical similarities between samples and to organize in clusters the calibration data base. The relationships between NIR spectra and modeled properties were well adapted for most of the prediction models in twice the interval of confidence at 95% of the reference methods after clustering of the data base into three clusters. Cluster analysis was necessary to improve the prediction quality of the PLS models.*

ABBREVIATIONS

CP	Principal Components
CV	Cross-Validation
DmodX	Euclidean distance from each point to the PLS model in the standardized predictors
DmodY	Euclidean distance from each point to the PLS model in standardized responses
FCC	Fluid Cracking Catalysis
IC	Interval of Confidence
LV	Number of PLS components selected by cross validation
Max lev.	Maximum leverage
MLR	Multiple Linear Regression
MS	Mass Spectrometry
NIR	Near InfraRed
PLS	Partial Least Squares regression
PRESS	Predicted REsidual Sum of Squares
RMSP	Root Mean Square error of Prediction
Total aromatics	Sum of the monoaromatics and the diaromatics+ compounds
UV	UltraViolet
X	Matrix of predictor variables
X-var	Predictor variation explained by the selected PLS factors
Y	Response variable
Y-var	Response variation explained by the selected PLS factors.

INTRODUCTION

Partial Least Squares (PLS) can be seen as a generalization of regression [1-6]. This method has a particular interest in chemometrics because, unlike Multiple Linear Regression (MLR), it can be applied for the analysis of data with strongly correlated (collinear) and/or noisy or numerous *X* variables (structural descriptors) and can simultaneously model several response variables *Y*.

Spectrometric calibration is a type of problem in which PLS regression can be very effective. The predictors are the spectral responses at different wavenumbers, and the responses are the amounts of various chemical properties in the sample.

In this work, the data consists of NIR recordings on 128 samples characterized by known concentrations of 11 chemical properties:

– the mol/100g of mono-, di+ and total aromatics determined by UV spectrometry;
– the wt% of mono-, di+ and total aromatics determined by UV spectrometry and mass spectrometry;
– the wt% of paraffins and naphthenes determined by mass spectrometry.

The samples were selected in order to cover the chemical diversity of the samples analyzed in the *IFP Research Center* during the last 10 years. So, as data analysis is based on an assumption of homogeneity, in this work, the calibration base was organized in three clusters using Ward's hierarchical clustering with Euclidean distance [7]. The characteristics of the different models for each cluster and for each reference method are discussed.

1 EXPERIMENTAL

1.1 Experimental Analytical Conditions

The NIR, MS and UV analysis of the 128 selected samples was performed on the same sampling along a period of 6 months in different laboratories.

1.1.1 Near Infrared Spectroscopy (NIR)

The near infrared spectra were recorded on a nitrogen purged Bomem MB160 spectrometer. It was equipped with a DTGS detector in transmission mode with a resolution of 4 cm^{-1} using a 2 +/– 0.02 mm cell (QX quality) and after a delay of 5 min with a dry nitrogen flow of 3l/min in the 4900-9200 cm^{-1} range. A maximum absorbance of around one absorbance unit was obtained in the wavenumber range 6400-4500 cm^{-1}. Each sample was measured twice randomly with 100 scans per spectrum. One of the two spectra was used for modeling if the spectral difference between the two recordings on the same sample was less than 0.002 absorbance units in the wavenumber range 6400-4500 cm^{-1}. The measurements were carried out at 27.5°C with the help of a Peltier cell [8] in a room where the temperature range could vary from 20°C to 30°C.

1.1.2 Mass Spectrometry (MS)

The analysis was performed on a mass spectrometer with double focalization MS50 manufactured by *Kratos*. The sample is introduced via a batch inlet heated at 270°C under a secondary vacuum. Analysis is performed by electronic impact at 70 eV and at medium resolution (R = 5000). Five spectra are acquired at 10 s/decade with a 41-302 uma (unit atomic mass) range. They are averaged in order to improve the signal to noise ratio. The wt% of monoaromatics, diaromatics+ (sum of the di-, tri- and polyaromatics), total aromatics (sum of mono- and di+), naphthenes and paraffins are determined following the method described in [9]. The precision of the method is given in Table 1.

1.1.3 UV Spectrometry

The analysis was performed on a UV spectrometer equipped with a double monochromator CARY4G manufactured by *Varian*. The sample, after dilution in cyclohexane,

is analyzed in transmission in a 0.2 mm quartz cell (QS quality). The mol/100 g and wt% of monoaromatics, diaromatics and triaromatics+ are determined following the method described in [10]. The total aromatics content was obtained by addition of the mono-, di- and polyaromatics. To calculate the w% of each family, the mean molecular mass of each family (*i.e.* monoaromatics, diaromatics and tri+) was determined by mass spectrometry for each aromatic family [9]. The precision of the reference method is given in Table 1.

TABLE 1

Interval of confidence (for one measurement) of the reference methods

	Interval of confidence (IC) for one measurement
MS method	C < 1%, IC = 0.2 C > 1%, IC = 3.54 10^{-2} C + 0.424 Quantification limit = 1%
UV method in mol/100 g	C > = 0.025%, IC = 2.12 10^{-2} C 0.0005 < C < 0.025% IC = 7.07 10^{-3} C + 3.54 10^{-4}
UV method in wt%	C > = 5%, IC = 2.12 C 0.1% < C < 5%, IC = 7.07 10^{-3} C + 7.07 10^{-2} Quantification limit = 1%

Interval of confidence at different levels of concentration of UV and MS methods for determination in wt%

Level of concentration in wt%	Interval of confidence by UV	Interval of confidence by MS
0.1	0.07	0.2 (< limit of quantification)
1	0.08	0.2
5	0.11	0.6
10	0.2	0.8
20	0.4	1.1
40	0.9	1.8

The intervals of confidence for one measurement calculated in Table 1 at different levels of concentration show that UV spectrometry analysis is much more reproducible (factor 3 approximately) than MS spectrometry.

2 CHARACTERISTTICS OF THE DATA BASE

2.1 Origin and Chemical Characteristics of the Samples

The data base includes kerosene, atmospheric and heavy atmospheric gas oil samples from numerous refining processes (*i.e.* FCC, direct distillation, coking and visbreaking and hydrotreatment). It mainly completes the data base described in reference [13] on which the wt% of hydrogen and cetane number were already modelled. Samples issued from hydrotreatment processes (Prime D, hydrocracking, H-oil) represent 65% of the database. Samples issued from FCC processes bring diversity to the paraffins/naphthenes/aromatics composition. They are more enriched in diaromatics than hydrotreated samples from the same distillation interval and so have a higher density.

Tables 2 to 6 illustrate the chemical ranges covered by the samples. The concentrations are determined in these tables by MS spectrometry.

TABLE 2

Chemical ranges covered by the 128 samples of the database

	Paraffins in wt%	Naphthenes in wt%	Monoaro. in wt%	Diaro+. in wt%	Total aro. in wt%
Min.	2.0	0	0	0	0
Max.	100	86.9	58.6	80.5	88.4

TABLE 3

Chemical ranges covered by the 12 samples of the data base issued from direct distillation

	Paraffins in wt%	Naphthenes in wt%	Monoaro. in wt%	Diaro+. in wt%	Total aro. in wt%
Min.	3	30.8	11	2.1	17.4
Max.	37.9	58.9	38.6	23.3	59.8

TABLE 4

Chemical ranges covered by the 22 samples of the data base issued from FCC

	Paraffins in wt%	Naphthenes in wt%	Monoaro. in wt%	Diaro+. in wt%	Total aro. in wt%
Min.	2	1.6	7.8	4.7	60
Max.	15.6	27.2	55.9	80.5	88.4

TABLE 5

Chemical ranges covered by the 2 samples of the database issued from thermal cracking

	Paraffins in wt%	Naphthenes in wt%	Monoaro. in wt%	Diaro+. in wt%	Total aro. in wt%
Min.	21.7	29.6	16.9	15.1	31.7
Max.	25.7	33.3	20	11.7	32

NB: olefins content are counted in aromatics.

TABLE 6

Chemical ranges covered by the 82 samples of the database issued from hydrotreatment processes

	Paraffins in wt%	Naphthenes in wt%	Monoaro. in wt%	Diaro+. in wt%	Total aro. in wt%
Min.	4.1	6.2	0.1	0.1	0.2
Max.	46.5	86.9	58.6	58.7	82.6

The range of the data base is illustrated by the following figures:

– Figure 1, which represents the mean distillation temperature in °C *versus* the density at 15°C;
– Figure 2, which represents the mean distillation temperature *versus* the wt% of total aromatics determined by MS;
– Figure 3, which represents the distribution of the different chemical families for the 128 samples of the data base;
– Figure 4, which represents the distribution of the different chemical families for each type of processes (*i.e.* direct distillation, FCC, thermal cracking and hydrotreatment). Samples with unknown origin are not represented.

2.2 Principal Components Analysis of the Calibration Base

The variation explained by the first principal components (PC), of the calibration base (predictors variables), is given by the Table 7.

TABLE 7

Variation explained by the first principal components

	CP1	CP2	CP3	CP4	CP5	CP6
Variation (%)	95.0	2.8	1.2	0.5	0.1	0.1
Total variation (%)	95.0	97.9	99.0	99.5	99.7	99.8

Table 7 indicates that only two or three principal components are required to explain almost all (97.9 and 99% respectively) of the variation in the calibration base (predictor variables). We can conclude that the *X* variables correlate well.

Figure 5 displays a plot of the calibration base projected onto the first three principal components. The plotting symbol is the process's type of all the samples in the calibration base. We notice that LCO GO issued from FCC processes are easily distinguishable and are on the right side on the PC2 axis.

2.3 Ward's Hierachical Clustering of the Data Base

A first PLS-1 modelisation of the properties on the complete database shows very bad statistical results *i.e.* RMSP of 0.02 (for example) for the prediction of the mol/100 g of monoaromatics by UV spectrometry. A more detailed exploration of the data by cluster analysis was thus necessary to improve the precision of the models.

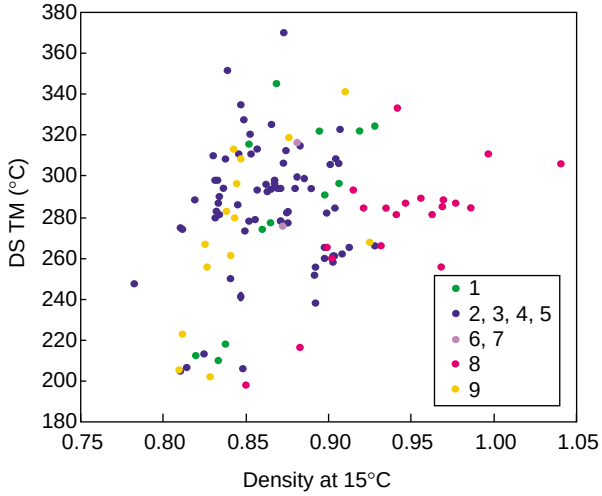

Figure 1

Mean distillation temperature *versus* density at 15°C.
In green: direct distillation, in blue: hydrotreatment, in violet: thermal cracking, in pink: FCC, in yellow: others.

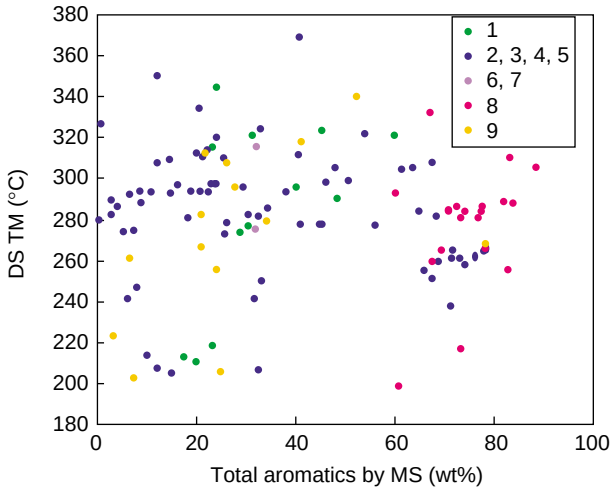

Figure 2

Mean distillation temperature *versus* the total aromatics content.
In green: direct distillation, in blue: hydrotreatment, in violet: thermal cracking, in pink: FCC, in yellow: others.

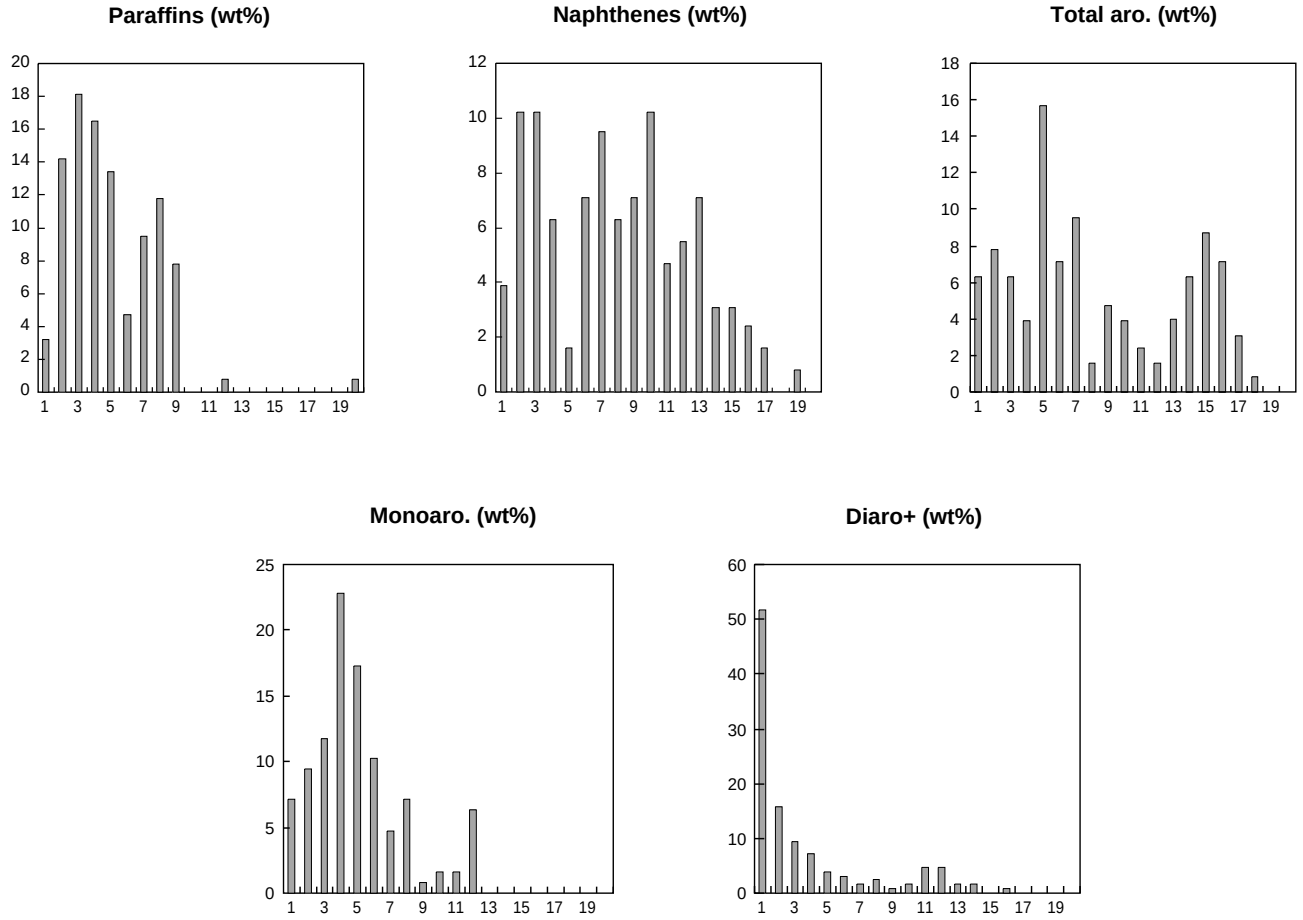

Figure 3

Distribution of the different chemical families for the 128 samples.

The abscissa axis represents the wt% of the considered family with a step of 5 wt% between 0 and 100% (20 classes). Class h: between $(h-1)\times 5$ and $h \times 5$ wt%, $h = 1$,..., 20.

To explore the various structures present in the calibration base, by taking into account their multidimensional character, the projection of the calibration base on the first ten principal components PC is explored by using Ward's hierarchical clustering method with the Euclidean distance. Figure 6 represents the dendrogram of this clustering with only the top 30 nodes. There are three main clusters in this base. The height of the nodes indicates the distance between the objects. The three clusters contain respectively 79, 24 and 25 objects. We can notice that cluster 1 (79 objects) and cluster 2 (24 objects) are close together when compared to cluster 3 (25 objects) and that the number of samples in two clusters (24 and 25 respectively) could be critical for robust modeling. Cluster 3 corresponds to samples issued mainly from FCC processes. Nevertheless, these two clusters cover a large range of PC1 and PC2 axis.

Figure 7 represents the calibration base plotted on its first three principal components. The labels are the clusters assigned by Ward's hierarchical clustering. This representation on the first three principal component's space displays the separation and the significance of the obtained clusters. The three clusters are very distinct on the projections on the three PC axis and they all describe a large variation of PC factors. By comparison with Figure 5, we can say that cluster 3 corresponds to samples with high aromatics and contains samples mainly issued from FCC processes.

Near infrared spectra of the samples of the clusters are represented in Figure 8. This representation shows the dissimilarity between the spectra in each of the three clusters and confirms the reliability required to keep the three clusters. The NIR fingerprints confirm the enrichment in paraffins for the cluster 1 and in aromatics for cluster 3.

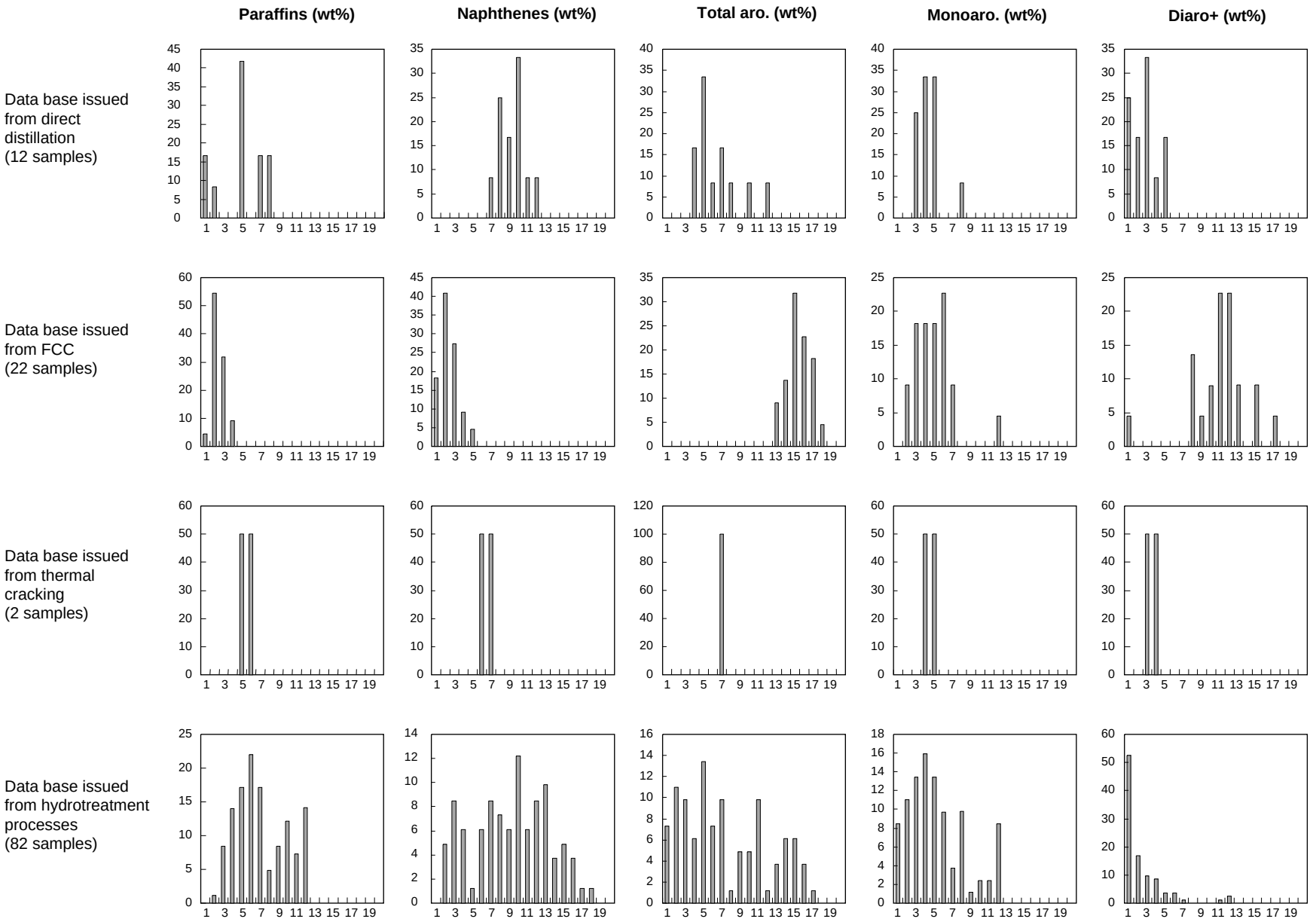

Figure 4

Distribution of the different chemical families for each type of processes. The classes are the same as on Figure 3.

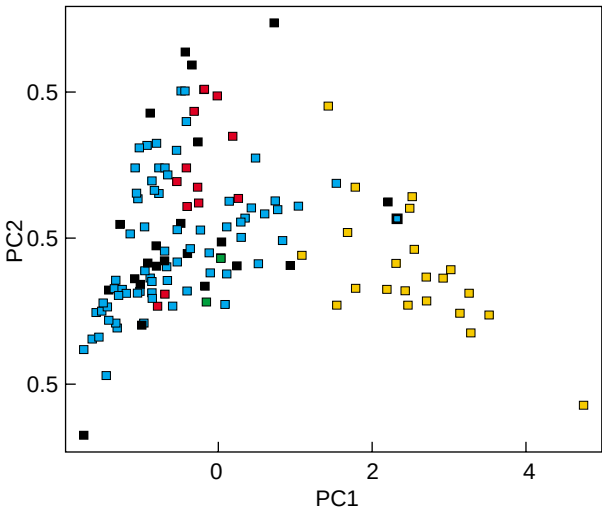

Figure 5

A plot of the calibration base projected onto the first two principal components.

The plotting symbol is the processes type of all the samples in the calibration base: FCC in red, direct distillation in yellow, thermal cracking in green, hydrotreatment in blue, others in black.

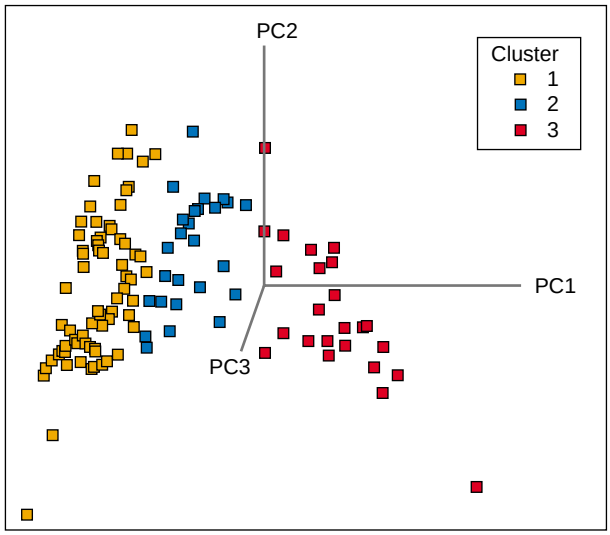

Figure 7

The calibration base plotted on its first three principal components.

Cluster 1 in yellow, cluster 2 in blue and cluster 3 in red.

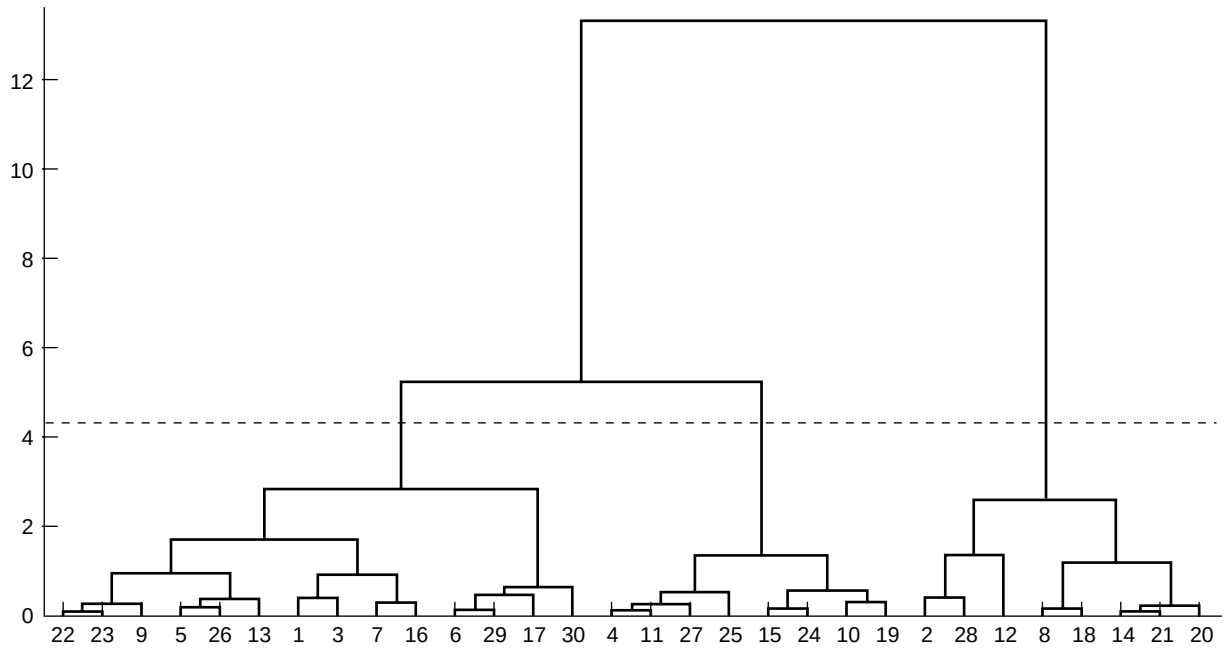

Figure 6

Dendrogram plot of the hierarchical cluster tree of the calibration base.

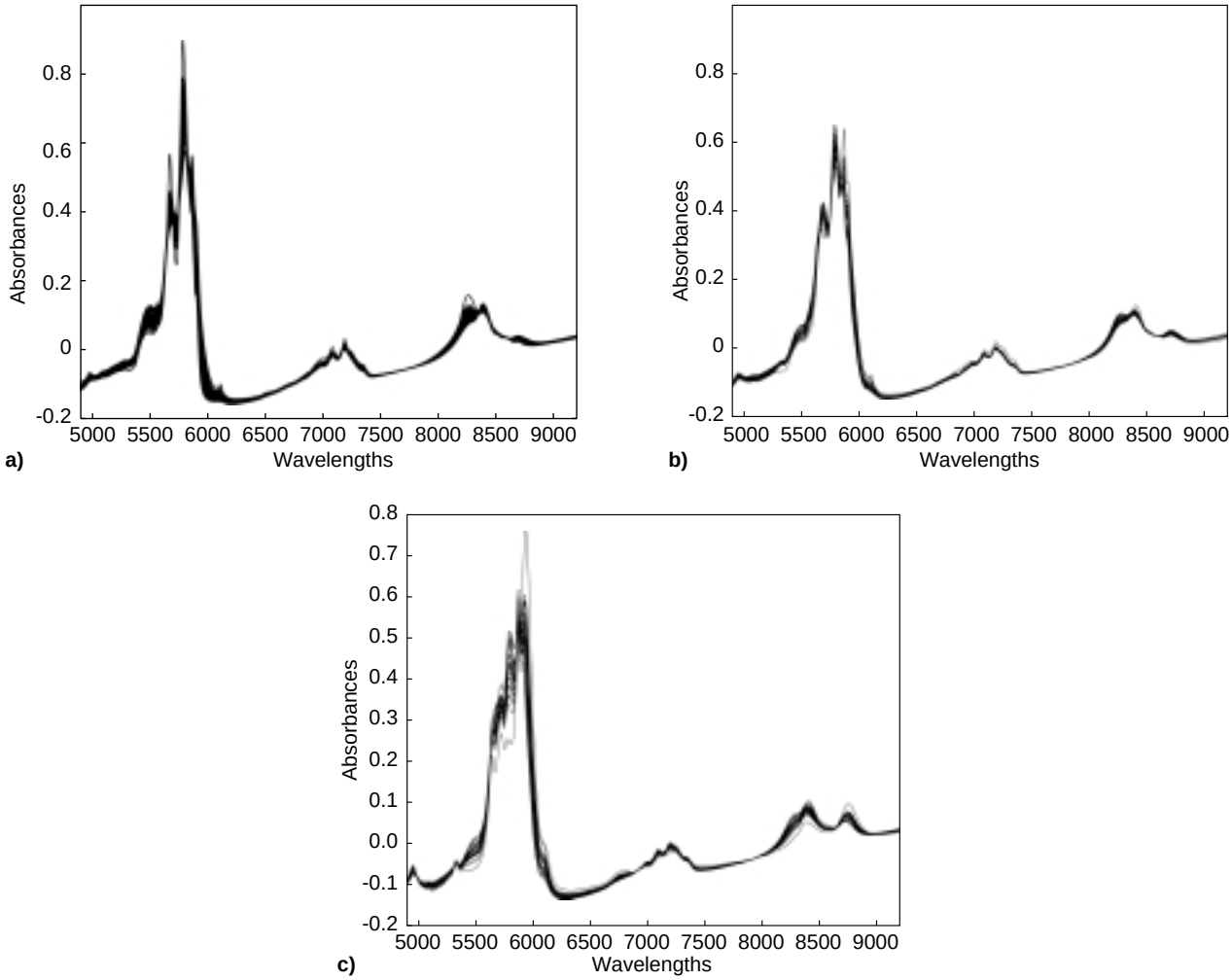

Figure 8

Near infrared spectra of the samples of the three clusters of the calibration base.
a) cluster 1; b) cluster 2; c) cluster 3.

In Figure 9 we present the box plot of the distributions of the properties to be modeled in each of the three clusters. These representations confirm the dissimilarity which exists between the three clusters as well as the correlation which exists between chemical properties and near infrared spectra.

Concerning the mono-, di+ and total aromatics, the shape of the three clusters depends neither on the reference method (UV or SM) nor on the unit (wt% or mol/100 g). The characteristics of the clusters are the following:

– Cluster 1 is characterized by the lowest content of total aromatics (at 95% in the 12-30 wt% range) and the higher content in paraffins and naphthenic compounds *i.e.* respectively at 95% in the range 17-25 wt% and 39-63 wt%. The content of monoaromatics (between 10-22 wt%) is greater than the diaromatics (at 95% in the range 0-7 wt%).
– Cluster 2 corresponds to an intermediate level of total aromatics (at 95% in the range 50-70 wt%), paraffins (at 95% in the range 15-23 wt%) and naphthenes (at 95% in the range 15-28 wt%). The amount of monoaromatics (at 95% in the range 25-60 wt%) is greater that the one of diaromtics+ (at 95% in the range 17-22 wt%).
– Cluster 3 corresponds to samples enriched in aromatics (total aromatics at 95% in the range 70-78 wt%) with a ratio monaromatics/diaromatic+ inversed by comparison with cluster 1 and cluster 2. The amount of paraffins and naphthenes are low. That corresponds to samples issued from FCC processes.

Finally, we compare the characteristics of the PLS model (for example) for the prediction of the mol/100 g of monoaromatics by UV spectrometry in the case of 1, 2 and 3 clusters. Clusters 1 and 2, were grouped in the case of

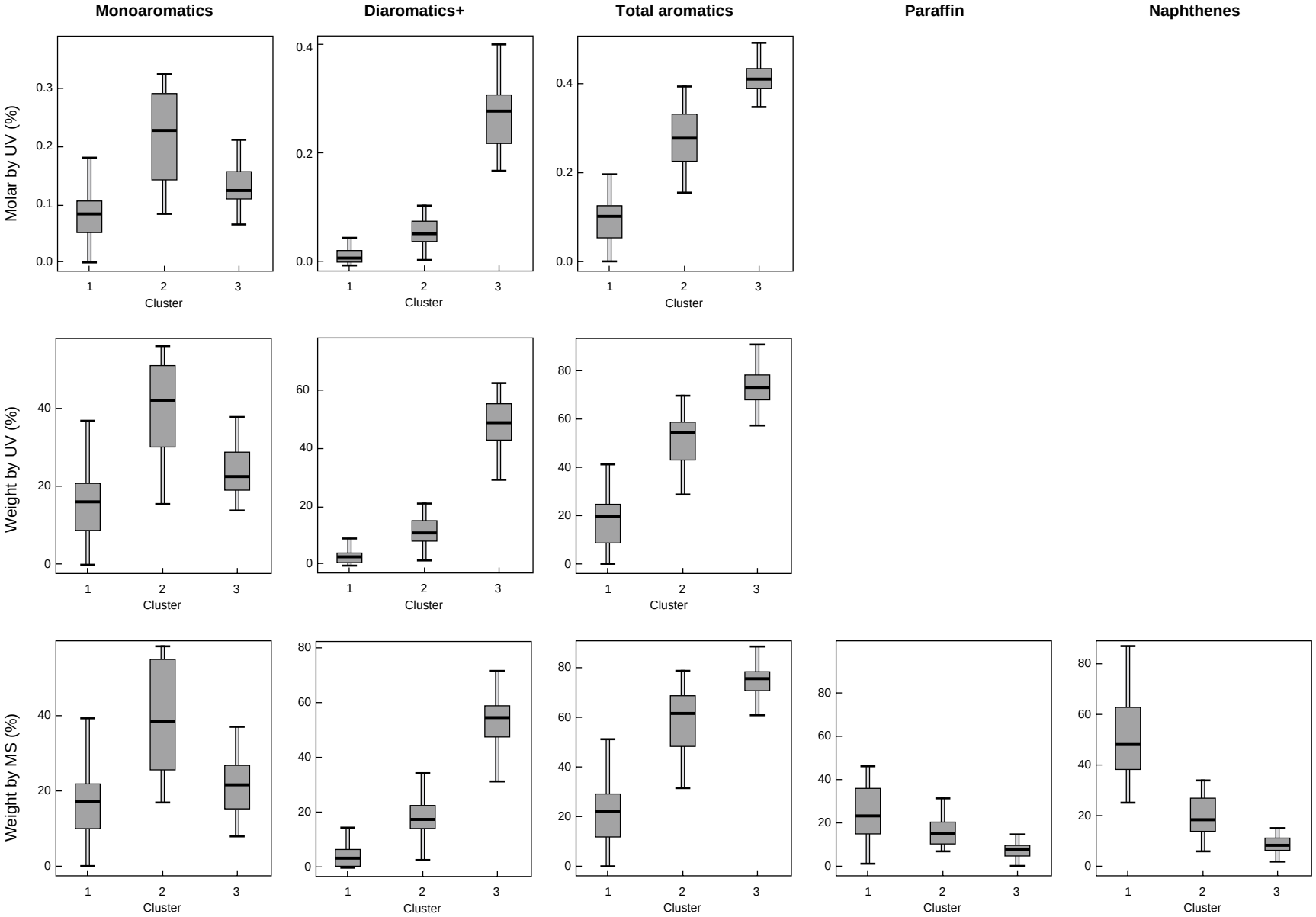

Figure 9

The box plot of the distributions of responses variables to be modeled in each of the three clusters.

two clusters. Table 8 shows the improvement of the NIR models by keeping the three clusters. So, to approach the precision of the reference methods, PLS-1 models were developed keeping the three clusters for every chemical property.

TABLE 8

RMSP of NIR model for the prediction of the mol/100 g of monoaromatics content by UV spectrometry, *versus* the number of clusters

Number of clusters	RMSP
1 cluster	0.02
2 clusters:	
cluster 1 + cluster 2	0.0126
cluster 3	0.0053
3 clusters:	
cluster 1	0.0059
cluster 2	0.0056
cluster 3	0.0053

3 STATISTICAL AND MATHEMATICAL TOOL USED FOR THE DEVELOPMENT OF THE MODELS

3.1 Partial Least Square Regression

PLS regression [11] is a recent technique that generalizes and combines features from principal component analysis (PCA) and multiple linear regression (MLR). It is particularly useful when we need to predict a set of response variables $(Y, n \times p)$ from a very large set of predictor variables $(X, n \times m)$.

The goal of the PLS regression is to predict Y from X and to describe their common structure. When Y is a vector and X is full rank, this goal could be accomplished using ordinary multiple regression (MLR). When the number m of the predictors is large compared to the number n of observations, matrix $X'X$ is likely to be singular and the regression approach is no longer feasible (*i.e.*, because of multicollinearity). Several approaches have been developed to cope with this problem. One approach called principal component regression is to perform a principal component analysis (PCA) of matrix X and then use the principal components to predict Y. The principal components are chosen to explain X rather than Y so nothing guarantees that they are relevant for Y.

PLS regression finds components from X that are also relevant for Y. Specifically, PLS regression searches for a set of components that performs a simultaneous decomposition of X and Y with the constraint that these components explain as much as possible of the covariance between X and Y. This step is followed by regression where the decomposition of X (PLS components) is used to predict Y. The "engine" of the PLS methodology is the nonlinear iterative partial least squares (NIPALS) algorithm [8]. Table 9 provides a summary of this algorithm.

TABLE 9

NIPALS algorithm for PLS regression

Step	Summary of steps	
0	Mean center X and Y:	$x_0 = X, y_0 = Y$
1	Calculate the $a^{\text{-th}}$ PLS factor	
2	Set the output score u_a equal to any column of y_{a-1}	
3	Compute input weights w_a by regressing x_{a-1} on u_a	$w_a = \frac{x'_{a-1}u_a}{u'_a u_a}$
4	Normalize w_a to unit length:	$w_a = \frac{w_a}{\|w_a\|}$
5	Compute the input scores t_a	$t_a = \frac{x_{a-1}w_a}{w'_a w_a}$
6	Compute output loadings q_a by regressing y_{a-1} on t_a	$q'_a = \frac{t'_a y_{a-1}}{t'_a t_a}$
7	Normalize q_a to unit length	$q_a = \frac{q_a}{\|q_a\|}$
8	Calculate new output scores u_a	$u_a = \frac{y_{a-1}q_a}{q'_a q_a}$
9	Check convergence on u_a: if yes go to step 10 else go to step 3	
10	Compute input loadings p_a by regressing x_{a-1} on t_a	$p'_a = \frac{t'_a x_{a-1}}{t'_a t_a}$
11	Compute inner model regression coefficient β_a	$\beta_a = \frac{t'_a u_a}{t'_a t_a}$
12	Calculate input residual matrix	$x_a = x_{a-1} - t_a p'_a$
13	Calculate output residual matrix	$y_a = y_{a-1} - \beta_a t_a \times q'_a$
14	If additional PLS dimension are necessary, replace x_{a-1} and y_{a-1} by x_a and y_a respectively and repeat steps 2 to 14	

When there is only one response variable, the standard PLS algorithm can be reduced to an algorithm referred to as PLS-1. In this work PLS-1 models were developed for every chemical property.

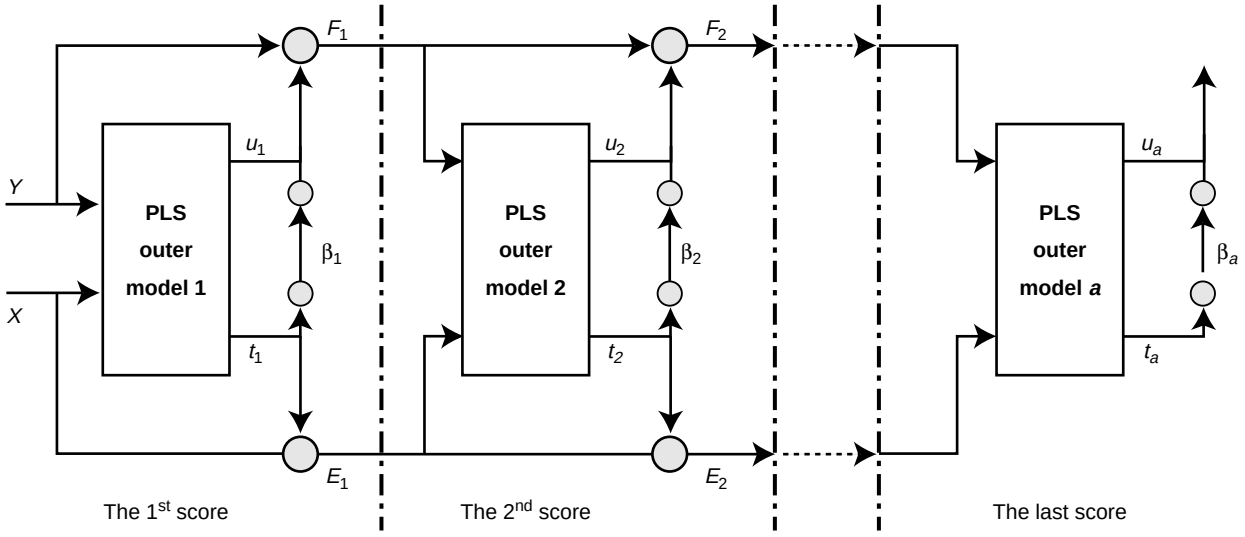

Figure 10

A schematic representation of the PLS method.

The power of PLS as a regression tool lies in the fact that it decomposes a multivariate regression problem into a number of uncorrelated univariate regressions *(Fig. 10).*

The partial least squares algorithm selects a number of orthogonal factors that maximize the covariance between each *X*-scores t_a and the corresponding *Y*-scores u_a. For a good PLS model, the first few factors show a high correlation between the *X*-scores and *Y*-scores. The correlation usually decreases from one factor to the next, whilst the higher order latent variables are typically associated with the random noise in the data. The appropriate number of latent variables can be chosen for example by means of cross-validation [12]. In this work, we use the *leave-one-out* cross validation presented below.

3.2 Leave-One-Out Cross Validation

Using a set of *n* calibration spectra, the PLS algorithm is performed on (*n*-1) calibration spectra and, with this calibration, the concentrations of the chemical properties of the sample left out during calibration is predicted. This procedure is repeated *n* times until each sample has been left out once. The prediction for each sample is then compared with the known value of the reference sample. The sum of the squared variable prediction errors for all calibration samples is a measure of how well a particular PLS model fits the response variable:

$$\text{PRESS} = \sum (Y_{\text{predicted}} - Y_{\text{measured}})^2$$

PRESS is calculated in the same way each time a new factor is added to the PLS model. The optimal order (number of factors) of the PLS model is the one that yields the minimum PRESS or root mean square error of Prediction RMSP:

$$\text{RMSP} = \sqrt{\frac{\text{PRESS}}{n}}$$

Data transformation: data are scaled to have a mean of zero.

3.3 Clustering

The clustering was performed by means of Ward's hierarchical clustering with Euclidean distance [7].

A PLS model effectively models both the predictors and the responses. In order to check and eliminate the outliers, in this work, we look at the Euclidean distance from each point to the PLS model in both the standardized predictors and the standardized responses (DModX and DModY). No point should be dramatically farther from the model than the rest.

3.4 Outlier Detection Tests

Whatever the algorithm used to correlate a signal (in this case the NIR spectrum) and property (the physio-chemical properties of gas oils, in this study), the PLS model, which correlates the signal and the property, can be applied in certain conditions (*i.e.* the chemical composition, the property range, the sample temperature). When one of these conditions is not

fulfilled, the unknown sample to analyze has to be declared as an "outlier".

In this work, two tests were used to detect outliers. The first was based on the analysis on the DModX and the second on the so-called "leverage value":

– The DModX of the unknown sample was evaluated and compared with the maximal DModX obtained by the calibration sample.
– The leverage value represents the fraction of variance explained by the sample. The leverage limit was generally set to the maximum one encountered with the calibration base. Sometimes, it was found necessary to take into account the density of points with the same level of leverage. The leverage limit was decreased to a value corresponding to a high-density population.

In this work, the calibration base was organized in three clusters; the unknown sample to analyze has to be classified to one of the three clusters on the base of their best fit to the respective model (DmodX, leverage).

4 PLS MODELS FOR THE PREDICTION OF THE AROMATICS FAMILIES

In this section, we will use the results of the performance of the PLS models as a tool to predict the concentrations of the aromatics families. As pointed out, the clustering and preprocessing of the outliers of raw data is a key step to obtaining a good data set for the calibration of the prediction models.

We will then present, in each of the three clusters, the characteristics of the prediction models: it will have three models for every property. The characteristics which will be presented are: the number *n* of objects (samples) used to elaborate the prediction models; the optimal number of PLS factors selected by cross validation (LV); max of the leverage (max. lev.); root mean square error of prediction (RMSP); predictor and response variation (*X-var* and *Y-var*) explained by the PLS factors selected for every model.

For each presented model, we will see that for clusters 2 and 3, the number of samples in each cluster is at the limit of

TABLE 10

The characteristics of the PLS models:
Mono-, di- + and total aromatics: mol/100 g by UV, wt% by UV and wt% by MS.
Naphthenes and paraffins: wt% by MS

Models	N	LV	Max. lev.	RMSP	(%) *X-var*	(%) *Y-var*	N	LV	Max. lev.	RMSP	*(%) X-var*	(%) *Y-var*	N	LV	Max. lev.	RMSP	(%) *X-var*	(%) *Y-var*
	Cluster 1						Cluster 2						Cluster 3					
Monoaro. by UV (mol/100 g)	68	6	0.99	0.0059	98.3	97.7	22	6	0.96	0.0056	98.9	99.1	20	6	0.91	0.0053	99.4	99.3
Monoaro. by UV (wt%)	68	7	0.53	1.02	98.6	98.6	20	6	0.72	0.90	97.9	99.5	19	6	0.76	0.88	99.0	99.1
Monoaro. by MS (wt%)	64	7	0.56	1.44	98.7	97.5	21	6	0.88	1.01	98.7	99.5	17	6	0.82	0.79	99.5	99.4
Di- + aro. by UV (mol/100 g)	69	6	0.99	0.0029	98.2	94.5	19	6	0.95	0.0027	99.1	99.4	21	6	0.98	0.0049	98.6	99.3
Di- + aro. by UV (wt%)	60	6	0.35	0.47	98.0	95.8	17	6	0.75	0.38	99.2	99.6	21	6	0.77	0.63	98.6	99.5
Di- + aro. by MS (wt%)	67	7	0.47	0.95	98.6	95.1	21	6	0.84	1.42	98.7	95.6	18	6	0.71	1.58	99.2	98.7
Total aro. by UV (mol/100g)	73	7	0.54	0.0031	98.5	99.6	21	6	0.89	0.0024	99.2	99.9	21	6	0.79	0.005	98.7	98.6
Total aro. by UV (wt%)	70	6	0.53	0.98	98.4	98.9	20	6	0.72	1.03	99.1	98.9	21	6	0.76	0.59	98.7	99.4
Total aro. by MS (wt%)	67	6	0.36	1.29	98.0	98.8	21	6	0.88	1.11	99.3	92.3	17	5	0.78	0.84	98.8	97.9
Naphthenes by MS (wt%)	72	6	0.68	2.07	96.2	98.3	21	6	0.89	0.99	99.3	97.6	16	4	0.67	0.69	97.1	95.3
Paraffins by MS (wt%)	72	6	0.65	1.62	97.4	98.7	19	5	0.89	0.72	99.1	99.5	17	4	0.72	0.73	97.4	93.5

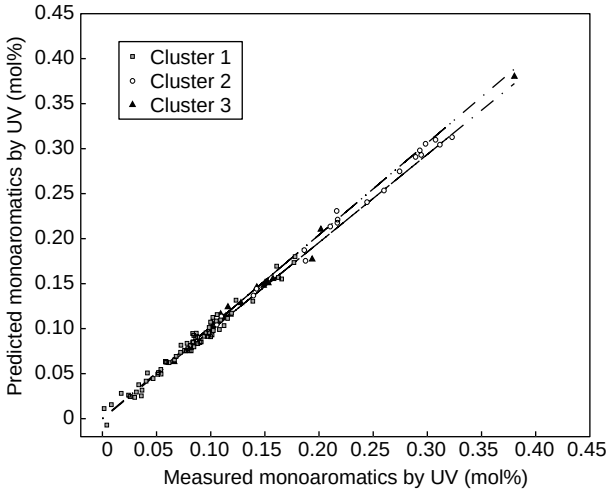

Figure 11

Predicted monoaromatics *versus* measured monoaromatics by UV in mol/100 g.

The interval of confidence is drawn with reference to the diagonal.

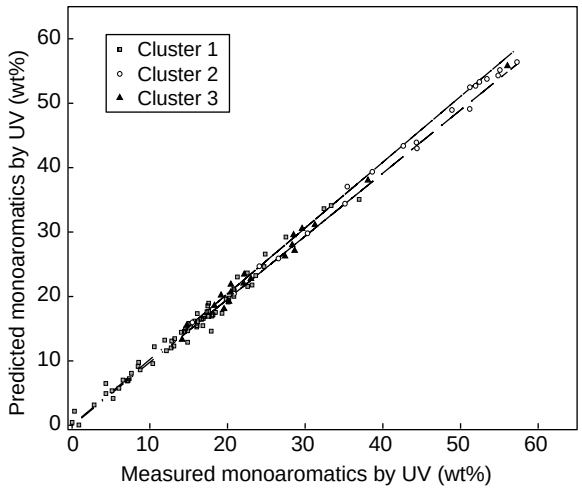

Figure 12

Predicted monoaromatics *versus* measured monoaromatics by UV in wt%.

The interval of confidence is drawn with reference to the diagonal.

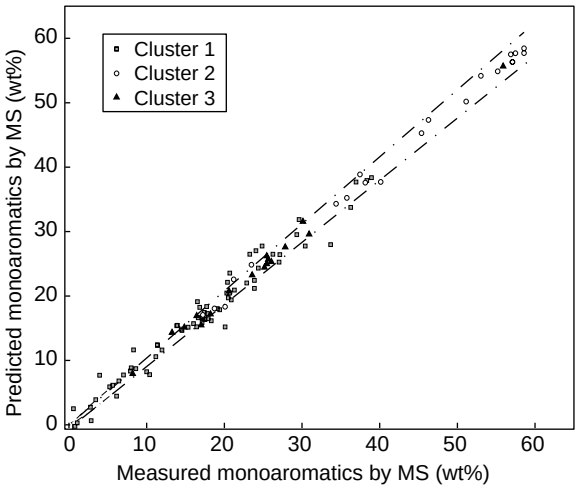

Figure 13

Predicted monoaromatics *versus* measured monoaromatics by MS in wt%.

The interval of confidence is drawn with reference to the diagonal.

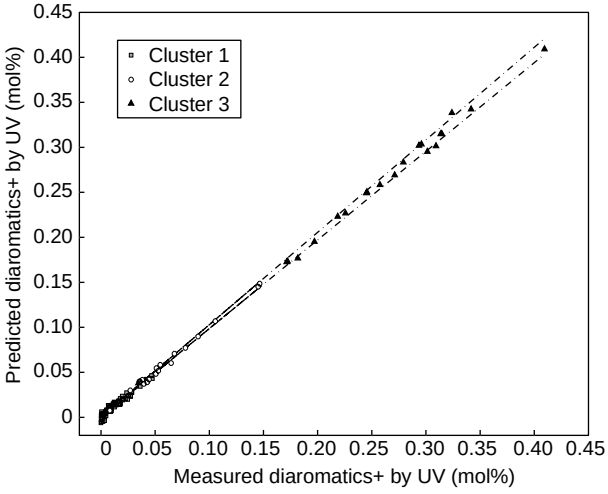

Figure 14

Predicted diaromatics+ *versus* measured diaromatics+ by UV in mol/100 g

The interval of confidence is drawn with reference to the diagonal.

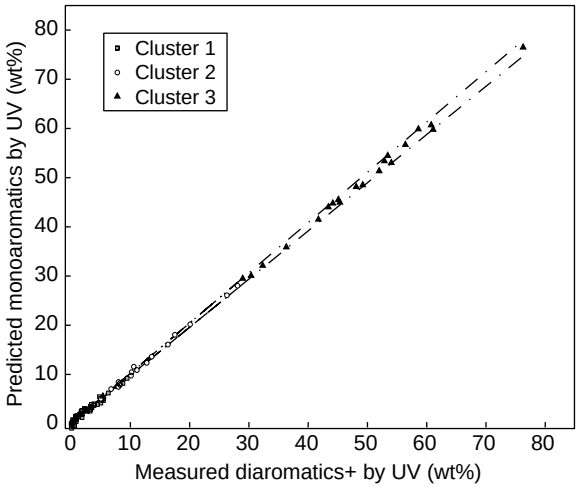

Figure 15

Predicted diaromatics+ *versus* measured diaromatics+ by UV in wt%.

The interval of confidence is drawn with reference to the diagonal.

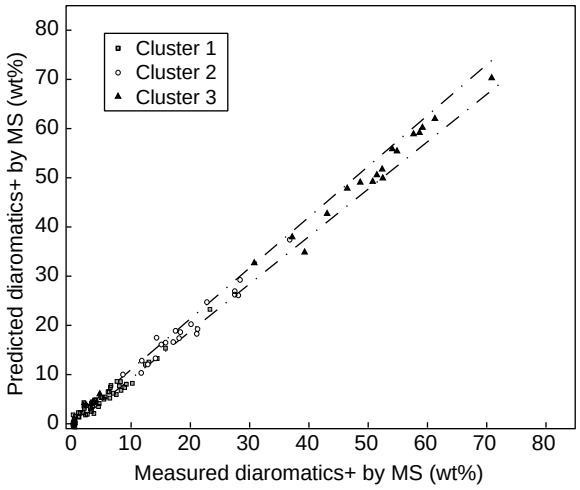

Figure 16

Predicted diaromatics+ *versus* measured diaromatics+ by MS in wt%.

The interval of confidence is drawn with reference to the diagonal.

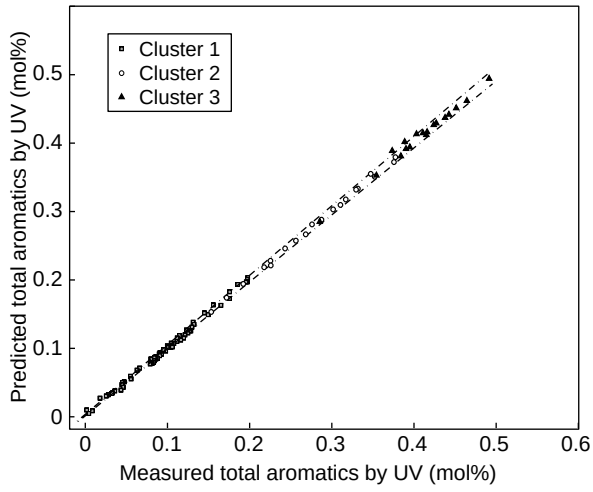

Figure 17

Predicted total aromatics *versus* measured total aromatics by UV in mol/100 g.

The interval of confidence is drawn with reference to the diagonal.

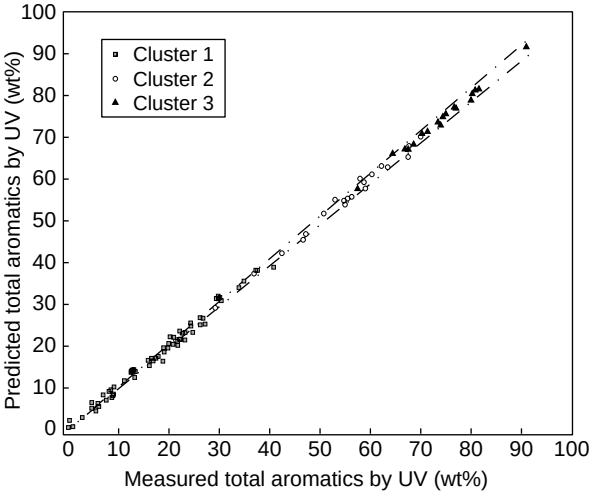

Figure 18

Predicted total aromatics *versus* measured total aromatics by UV in wt%.

The interval of confidence is drawn with reference to the diagonal.

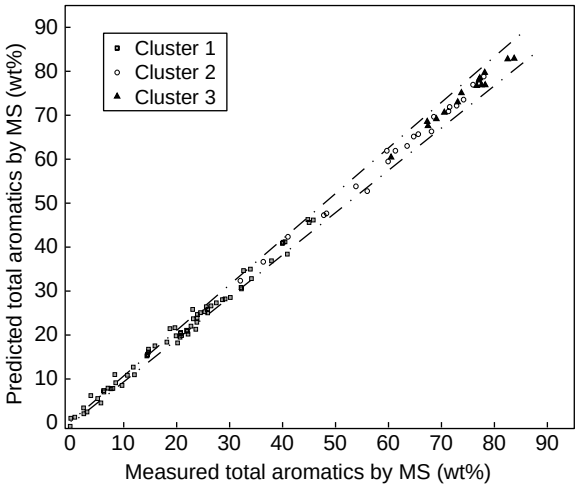

Figure 19

Predicted total aromatics+ *versus* measured total aromatics by MS in wt%.

The interval of confidence is drawn with reference to the diagonal.

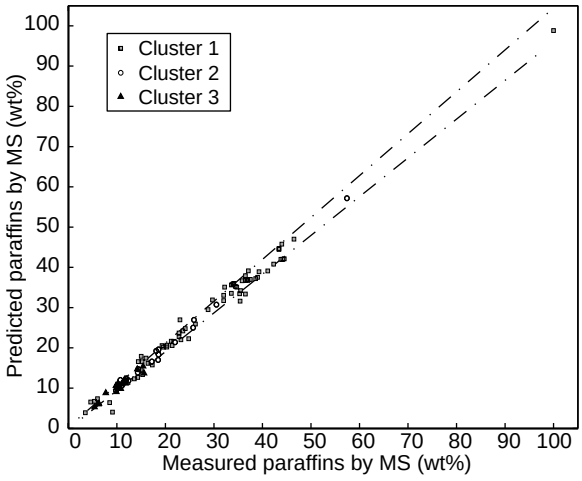

Figure 20

Predicted paraffins *versus* measured paraffins by MS in wt%.

The interval of confidence is drawn with reference to the diagonal.

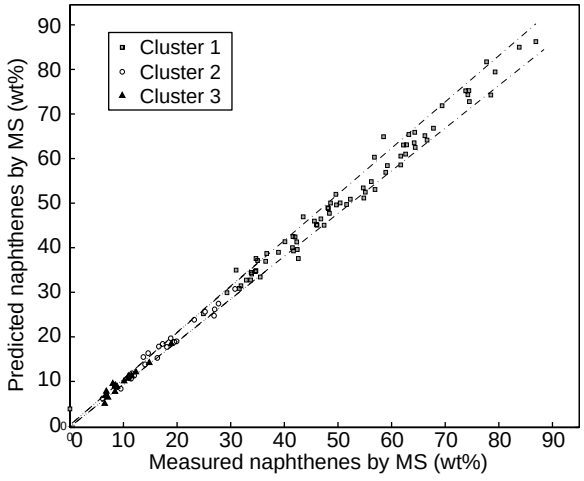

Figure 21

Predicted naphthenes *versus* measured naphthenes by MS in wt%.

The interval of confidence is drawn with reference to the diagonal.

acceptance considering the number of factors. For the latter, 95% of the predicted values are in the interval of confidence of the reference method. That confirms the reliability of the clustering in three clusters.

In the Table 10 we present for each of the three clusters the characteristics of the prediction models of the properties, mol/100 g by UV, wt% by UV and wt% by MS of monoaromatics, diaromatics+ and total aromatics. We also present the characteristics of the prediction models of the wt% by MS of naphthenes and paraffins. Outlier detection decreases the number of samples in each determined cluster.

The variation summary of the prediction models shows that over 97% of the predictor variation and over 92% of the response variation are accounted for by the PLS components selected by cross validation. They show the explanatory power of the models and a correlation between chemical properties and near infrared spectra.

Graphs below (Figures 11-21) show in each of the three clusters the quality of the prediction of the NIR models of aromatics families. They allow us to describe the global relation between the response variables (aromatics families) and the NIR spectra.

5 ANALYSIS OF THE PERFORMANCES OF THE NIR MODELS

Figures 22 and 23 represent, for mono-, di+ and total aromatics, the correlation between the UV and MS determinations (in wt%) respectively between the reference data and the NIR predictions. We can say that NIR modeling keeps the correlation between the two reference methods for the determination of aromatic families.

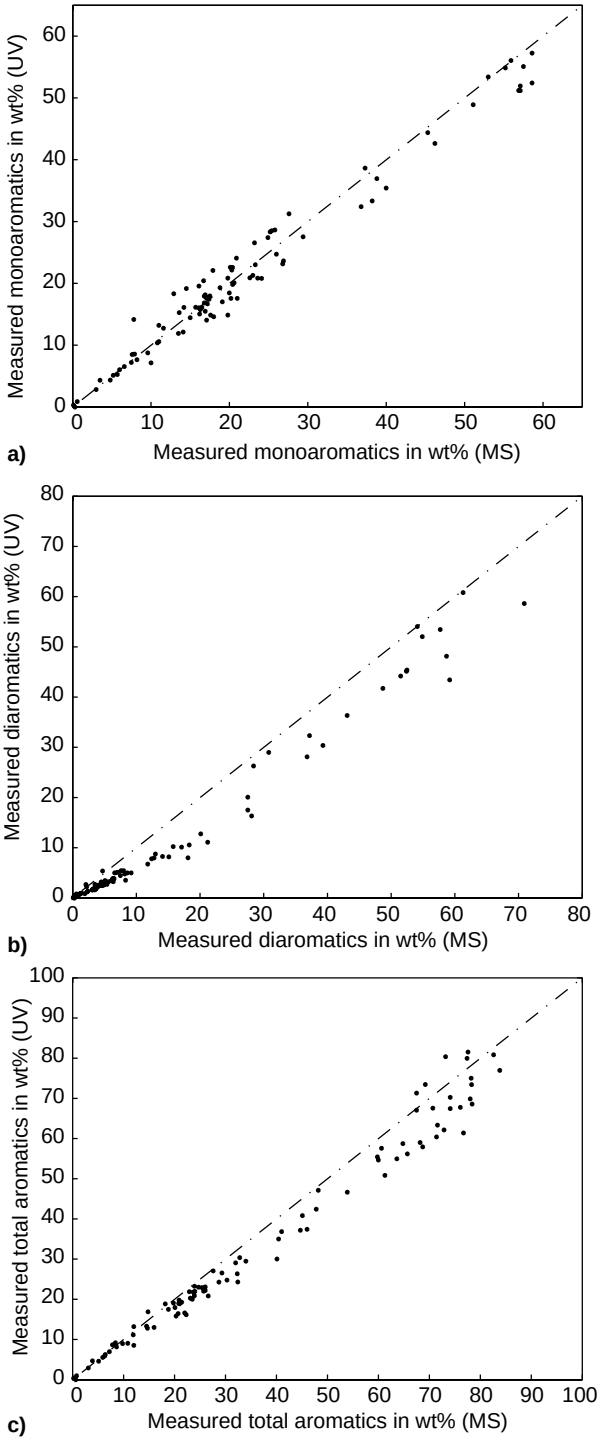

Figure 22

Correlation between UV and MS results with the references values.

a) measured wt% monoaromatics by UV *versus* measured wt% monoaromatics by MS.

b) measured wt% diaromatics+ by UV *versus* measured wt% diaromatics+ by MS.

c) measured wt% total aromatics by UV *versus* measured wt% total aromatics by MS.

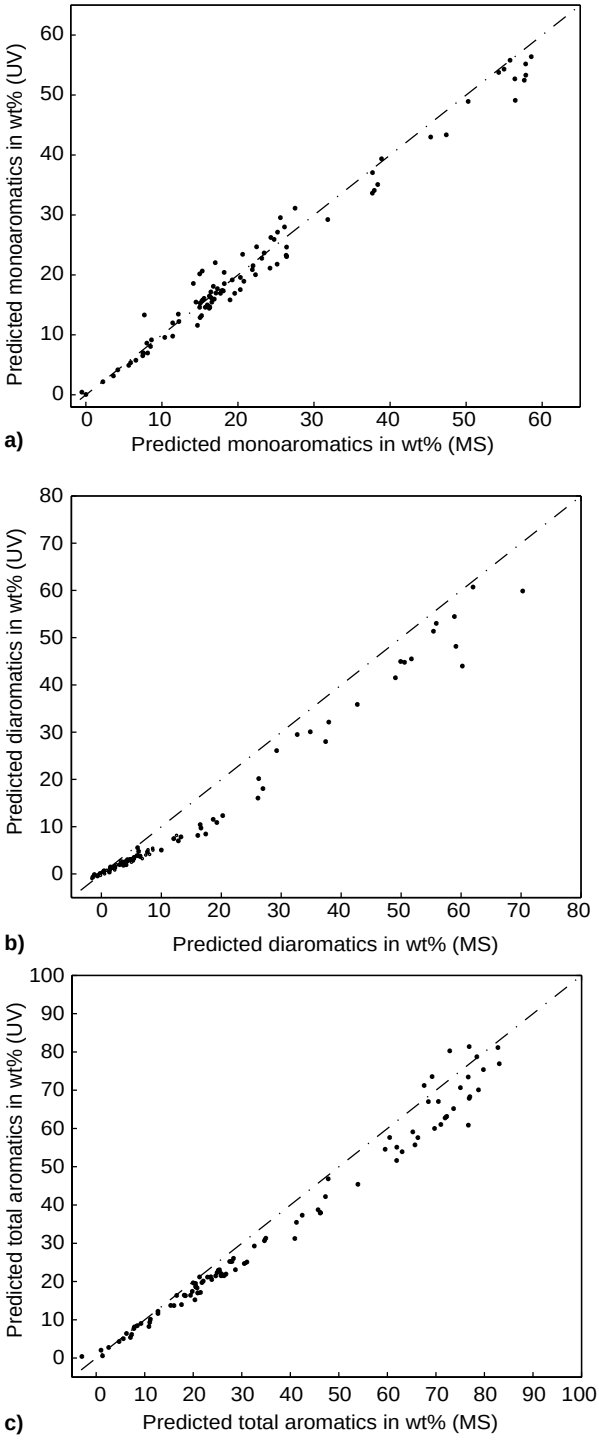

Figure 23

Correlation between UV and MS results with NIR predicted values.

a) NIR predicted wt% monoaromatics by UV *versus* NIR predicted wt% monoaromatics by MS.

b) NIR predicted wt% diaromatics+ by UV *versus* NIR predicted wt% diaromatics+ by MS.

c) NIR predicted wt% total aromatics by UV *versus* NIR predicted wt% total aromatics by MS.

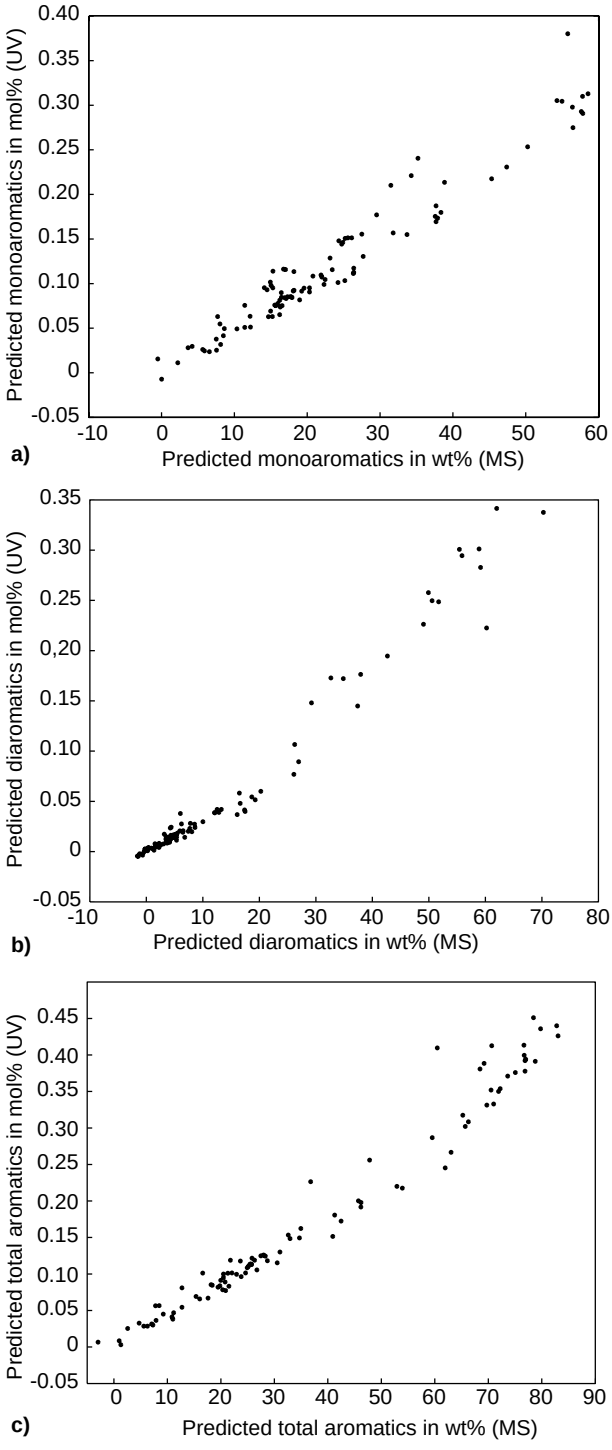

Figure 24

Effect of the determination of *M* for each class of aromatics by mass spectrometry on the correlation between UV and MS method.

a) for monoaromatics, b) on diaromatics+, c) on total aromatics.

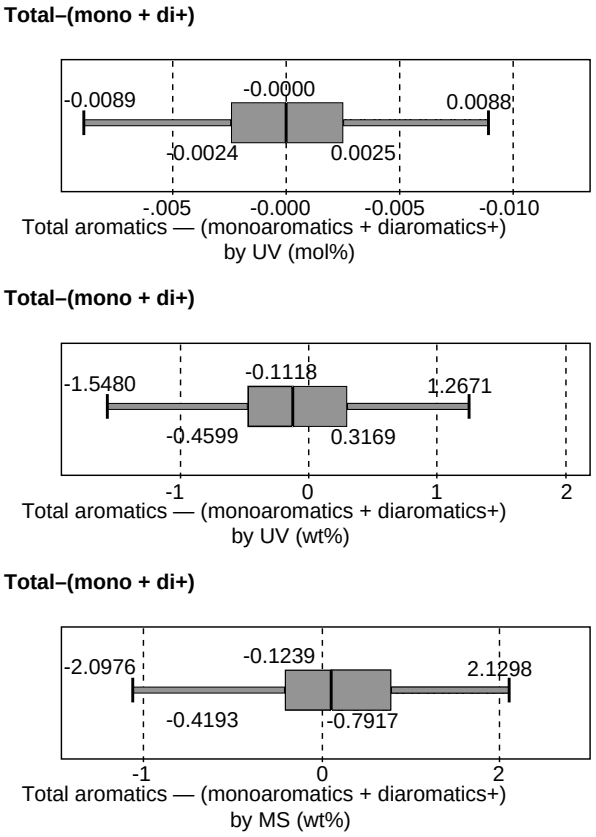

Figure 25

Differences between NIR total aromatics predictions and (mono + di+) NIR predictions.

Comparing each family of aromatics and each cluster, the correlations between the mol/100 g by UV *versus* the wt% by MS *(Fig. 24)* and between the wt% by UV *versus* the wt% by MS *(Fig. 23)*, we can say that the introduction of the molecular mass for each family of aromatics improves the correlation. So the developed NIR models referring to the UV method in wt% is interesting. Usually in a laboratory, it is costly to determine *M* for each type of aromatic. Because of this, *M* is usually considered equal for each family of aromatics and equal to that of the sample and calculated by correlation. So, the wt% of each family of aromatics determined by UV is directly proportional to the mol/100 g determined by UV.

Figure 25 illustrates for each determination in wt% by UV and MS and in mol/100 g by UV, the differences (total – (mono+ di+)). We can see that the differences are centred around 0 with 95% of the differences very close to 0. That illustrates the coherence of the NIR models.

Table 11 gathers the performances of the NIR models regarding the % of predicted samples within the 2σ range

(resp. within 4σ) of the reference method. Concerning the models referring to the MS method, the mono- and total aromatics, naphthenes and paraffins models are acceptable due to the fact that 95% of the samples are within 4s of the reference method. For the diaro.+ model, 90% of the NIR prediction are within the 4σ range of the reference method due to the low concentration of diromatics+.

Concerning the UV models, they seem to be less performant than the MS one but the UV method is three times more reproducible than the MS method. So, the statistics can be evaluated as being satisfactory for process control. The NIR predictions are very reproducible [14] and samples will be able to be compared as long as they belong to the same series.

TABLE 11

Percentage of predicted samples of the calibration data base in the 2σ (resp in 4σ) of the reference methods

	Mono-		Di+		Total		Napht.		Paraff.	
	2σ	4σ	2σ	4σ	2σ	4σ	2σ	4σ	2σ	4σ
UV mol/100 g	45	81	34	61	72	97				
UV wt%	46	83	37	73	57	87				
MS wt%	68	94	50	90	76	96	75	96	71	98

CONCLUSION

This work has shown that on a large chemical diversity of samples, the NIR spectrum of mid-distillates contain the chemical information to model by the PLS algorithm different chemical families such as paraffins, naphthenes and total aromatics. Furthermore, it is possible to quantify the monoaromatics and diaromatics+.

We have demonstrated that:

– the NIR models keep the correlation between UV and MS data from the reference methods;
– the coherence of the three models total aromatics, monoaromatics and diaromatics+;
– the introduction of the mean molecular mass of each aromatics family improves the correlation between UV and SM data. NIR models will then present a certain advantage in practice because it is very costly to determine *M* by MS;
– the NIR prediction of the wt% paraffins, naphthenes, mono- and total aromatics are for 95% within the 4σ range of the reference method, which is satisfactory. Only 90% of the NIR predictions are within the 4σ range of the reference method for the model "diaromatics+". This is due to the lower concentrations.
– taking into account that the UV method is three times more reproducible than MS method and that NIR predictions are more stable than UV method, the UV models could be used for process control to compare samples from the same series with high precision.

The PLS algorithm has demonstrated its limitation in modeling a large variety of samples and clustering, which groups samples with chemical similarities, was necessary. Other algorithms such as topology, PLS-2 or neural networks could be tested in the future on this large data base. Calculation of the errors on the NIR predictions with algorithms such as bootstrap could be interesting in our case and will soon be implemented.

This work has completed the panel of properties, it is possible now to predict off-line, in-line and on-line at the *IFP Research Centre*. Several properties of interest for the characterization of mid-distillates *i.e.* cetane number, wt% hydrogen, wt% and mol/100 g mono-, di+ and total aromtics, naphthenes and paraffins can now be determined simultaneously, without any delay, on very small amounts of sample if necessary.

Even if the data base already covers a wide range of chemical compositions, we are always adding new samples to the data base by constant evaluation of them.

REFERENCES

1 Hoskuldsson, A. (1996) *Prediction Methods in Science and Technology*, **1**, Basic Theory, Thor Publishing, Denmark.

2 Martens, H. and Naes, T. (1989) *Multivariate Calibration*, John Wiley & Sons, Chichester.

3 Stahle, L., and Wold, S. (1988) *Multivariate Data Analysis and Experimental Design in Biomedical Research. Progress in Medical Chem.*, **25**, Ellis, G.P., and West, G.B., eds., Elsevier Science.

4 Tenenhaus, M.(1998) *La régression PLS : Théorie et pratique*, Éditions Technip.

5 Wold, H. (1982) Soft Modeling. The Basic Design and Some Extensions. In: *Systems under Indirect Observation*, II, Jöreskog, K.G., and Wold, H., eds., North-Holland, Amsterdam.

6 Wold, S., Johansson, E., and Cocchi, M. (1993). PLS - Partial Least Squares Projections to Latente Structures. In*: 3D QSAR in Drug Design, Theory, Methods, and Applications*, Kubinyi, H., ed., ESCOM Science Publishers, Leiden.

7 Saporta, G. (1990) *Probabilités, statistique et analyse des données*, Éditions Technip, 488.

8 Hoffmann, U. and Zanier-Szydlowski N. (1999) Portability of near infrared spectroscopic calibrations for petrochemical parameters, *Journal of Near Infrared Spectroscopy*, **7**, 33-45.

9 Castex, H., Boulet, R., Juguin, J. and Lepinasse, A. (1983) Analyse des kérosènes et des gazoles moyens par spectrométrie de masse à moyenne résolution. *Revue de l'Insitut Français du Pétrole*, Éditions Technip, **38**, 4.

10 Burdett, R.A., Taylor, L.W., Jones, L.C. (1955) Determination of aromatic hydrocarbons in lubricating-oil

fractions by far UV spectroscopy. *Journal of Molecular Spectroscopy*, Report of a Conference, 1954, 30-41.

11 Geladi, P. and Kowalski, B.R. (1986) Partial least-squares regression - a tutorial. *Analytica Chimica Acta*, **185**, 1-7.

12 Wold, S. (1978) Cross-validatory estimation of the number of component analysis. *Technometrics* **20,** 4, 397-405.

13 Zanier-Szydlowski, N., Quignard, A., Baco, F., Biguerd, H., Carpot, L and Wahl, F. (1999) Control of refining processes on mid-distillates by near infrared spectroscopy. *Oil & Gas Science and Technology, Rev. IFP*, **54**, 4, 463-472.

14 Zanier-Szydlowski, N., Berger, M., Wahl, F. and Guillaume, D. (2003) Performance of a near infrared spectrometer equipped with an autosampling accessory. *Journal of Near Infrared Spectroscopy*, 11, 83-95.

Final manuscript received in March 2004